SOLUTION BRIEF

Rechenzentrum
High Performance Computing

Optimal gerüstet für die Konvergenz von HPC und KI

Bestandsaufnahme: Wie geeignet ist Ihre HPC-Plattform für effiziente KI-getriebene Workloads

intel select solution

Kurzübersicht

Auf dem Gebiet der künstlichen Intelligenz (KI) entwickeln sich allgemeines Leistungsvermögen, Datenvolumen und Rechenleistung stetig weiter. KI ist daher inzwischen für viele Unternehmen eine ernst zu nehmende Option. In Kombination mit High-Performance-Computing (HPC) steigert sich ihr Potential sogar noch weiter. Beide Ansätze zusammenzuführen erfordert jedoch sorgfältige Überlegung und Planung. Zunächst einmal müssen neue KI-Maßnahmen im Einklang mit der Unternehmensstrategie stehen. Dann müssen sich die KI-Workloads in Ihre bestehende HPC-Infrastruktur einfügen, um auf möglichst effiziente und kostengünstige Weise bestmögliche Ergebnisse zu erzielen. Dieses Whitepaper skizziert die grundlegenden Überlegungen, die Unternehmen anstellen müssen, die KI in ihre HPC-Umgebung integrieren möchten. Außerdem werden konkrete erste Schritte aufgezeigt, mit denen die Konvergenz von HPC und KI erfolgreich vorangetrieben werden kann.

Grundsätzliche Überlegungen zur Konvergenz von KI und HPC

Für HPC-Profis bedeutet der Vormarsch der KI Herausforderung und Chance zugleich. Obwohl der Einsatz von KI bereits seit Jahrzehnten erforscht wird, reichten die Menge der verfügbaren Daten und die Rechenleistung lange Zeit nicht aus, um Erkenntnisse gewinnen und tragfähige Entscheidungen treffen zu können. Anders heute: Die meisten wissenschaftlichen Einrichtungen, Regierungen und Unternehmen generieren eine wahre Datenflut, parallel dazu steigt die Rechenleistung und die Kosten sinken. Dadurch wird KI endlich eine gangbare Option.

Als Folge dieser Entwicklung wird heute viel darüber diskutiert, inwieweit KI-Lösungen – durch Konvergenz mit bestehenden HPC-Workloads wie Simulation und Modellierung – die Unternehmenswelt verändern und den Erkenntnisgewinn in Forschung und Wissenschaft beschleunigen werden.

Ganz so einfach ist es allerdings nicht. Zunächst einmal muss geklärt werden, was sich hinter dem Begriff Konvergenz verbirgt. Eine einheitliche Definition gibt es nicht. Je nach Bedarf und Anwendungsfall können KI und HPC ganz unterschiedlich kombiniert werden. Die folgenden Szenarien beispielsweise könnten alle als Konvergenz von HPC und KI bezeichnet werden:

- Die Implementierung und Ausführung von KI-Frameworks wie Tensorflow* auf einer bestehenden HPC-Infrastruktur.
- Die Optimierung von bestehenden HPC-Workloads (z. B. Simulation, Modellierung) durch den Einsatz einer KI-Engine, die die Ausgabedaten nach dem Durchlauf eines Modells analysiert.

- Der Einsatz von Generative Adversarial Networks (eine Klasse von KI-Algorithmen für unüberwachtes maschinelles Lernen) zum Kombinieren komplexer Datenquellen. Nehmen wir an, im Rahmen einer kosmologischen Studie zu Dunkler Materie wird eine lineare algebraische Gleichung auf einem HPC-Cluster ausgeführt, um statistische Modelle zu generieren. Durch das Hinzufügen einer KI-Schicht auf der gleichen Plattform können zusätzlich Daten und Bilder von Satelliten eingebunden werden, um schnellere Ergebnisse zu erzielen und die Modelle zu optimieren.

Diese Beispiele sind erst der Anfang. Um Workloads derart verschmelzen zu können, ist es wichtig, auch die zugrunde liegenden Technologie-Stacks im Blick zu behalten. Während sich die Anforderungen auf Hardware-Ebene teilweise überschneiden können, sind die benötigten Programmier- und Software-Stacks sehr unterschiedlich. HPC-Anwendungen laufen normalerweise auf einer Bare-Metal-High-End-Infrastruktur mit optimierter Middleware (z. B. parallele Dateisysteme, optimierte und Domain-spezifische HPC-Bibliotheken). KI-Anwendungen dagegen laufen heute in der Regel auf Public- oder Private-Cloud-zentrierter Infrastruktur und Middleware wie Hadoop Distributed File System* (HDFS*), Spark* und S3*.

Ebenso wie sich Konvergenz aus ganz unterschiedlichen Konstellationen ergeben kann, gibt es auch nicht den einen Stack, der für alles passt. Wer heute KI und HPC in seiner eigenen Umgebung zusammenführen will, der muss zwei wesentliche Dinge beachten:

- Trotz des Hypes ist KI nicht zwangsläufig die Lösung für alles. Es ist wichtig mit den strategischen Herausforderungen und Prioriäten Ihres Unternehmens zu starten und festzustellen ob und - falls ja - wo KI diese unterstützen kann. Wenn Sie KI nur um ihrer selbst willen und ohne strategischen Nutzen einführen, bringt Ihnen das wenig.
- Die Konvergenz von HPC und KI hat bereits begonnen, die Best Practices entwickeln sich weiter. Auch Sie können schon heute den ersten Schritt in Richtung Konvergenz machen, indem Sie Ihre bestehende Infrastruktur als Ausgangspunkt dafür nehmen.

Einstieg

Führen Sie sich zunächst Ihre Unternehmens- und Projektziele vor Augen und klären Sie ab, wo Sie mit KI schnellere, kostengünstigere und präzisere HPC-Ergebnisse erzielen können. Auf dem Gebiet der KI ist vieles noch in Entwicklung. Forschungsinstitute, Universitäten und Unternehmen sind noch dabei auszuloten, wie sie dank KI schneller ans Ziel kommen. Wenn Sie gerade erst begonnen haben, KI zu nutzen oder zu entwickeln, sollten Ihre Investitionen in erster Linie darauf abzielen, die Wertschöpfung und/oder die Produktivität Ihrer bestehenden Infrastruktur zu erhöhen. Parallel dazu sollten Sie Ihre Algorithmen entwickeln und optimieren und Ihre Modelle trainieren, um Sie auf die Bedürfnisse des Unternehmens abzustimmen. In den meisten HPC-Rechenzentren ist die bereits vorhandene Ressource ein auf Intel® Xeon® Prozessoren basierender Cluster, der für eine Vielzahl von KI-Workloads optimiert werden kann. Es ist nicht notwendig, sofort in neue und kostenintensive Beschleuniger und/oder GPUs zu investieren. Das sollten Sie erst dann tun, wenn Sie sich darüber im Klaren sind, ob Sie sie wirklich brauchen und wie Sie sie effizient nutzen können. Wenn Sie sich unsicher sind, welche Anwendungsfälle Sie zuerst angehen sollten, sehen Sie sich an, wie Andere den Einstieg geschafft haben.

Seien Sie sich auch darüber im Klaren, dass KI heute keine schlüsselfertige Lösung ist. Das Entwickeln hochwertiger Trainingsmodelle erfordert ein spezifisches Set an Entwicklerkompetenzen und -ressourcen; ebenso das Implementieren dieser Modelle in Ihre HPC-Prozesse, um daraus Erkenntnisse zu gewinnen. Die Datenanalysten, die in der Regel für den Einsatz von KI verantwortlich sind, verfügen meist nicht über die Programmierkenntisse, um KI-Frameworks zu verwenden. Daher müssen sie von Personen geschult oder unterstützt werden, die über diese Kompetenz verfügen. Das ist ebenso wichtig wie die Wahl der richtigen Technologie. Diese Kompetenz können Sie durch regelmäßige Weiterbildungen und Schulungen der Intel® AI Academy erweitern und vertiefen. Die Intel® AI Academy gibt Ihnen Hilfestellung und Tipps dazu, wie Sie die Intel® Technologie für Deep Learning und andere KI-Workloads nutzen können, egal wo – von der Cloud bis zum Netzwerkrand.

Praxisbeispiel

Hirnforschung an der Princeton University

Das Institut für Neurowissenschaften der Princeton University hat in Zusammenarbeit mit Intel Labs eine Software entwickelt, die das menschliche Gehirn in Echtzeit abbildet. Mit Hilfe von HPC, maschinellem Lernen und künstlicher Intelligenz analysiert sie Daten von Bildern, die mittels funktioneller Magnetresonanztomografie (fMRT) gemacht wurden, und deckt so auf, was im Gehirn passiert. Das Team hat ein Kit zur Analyse von Gehirnaufnahmen entwickelt, mit dem jetzt Wissenschaftler in aller Welt an den Forschungen teilhaben können. Die Ergebnisse dienen zur Verbesserung der Diagnose und Behandlung psychiatrischer Störungen.

Fünf Schritte zur Optimierung

Intel arbeitet mit Kunden in Wissenschaftsbetrieben, Regierungen und Unternehmen zusammen und unterstützt sie dabei, die Implementierung von KI-Funktionalität in ihre HPC-Umgebung zu prüfen, zu planen und durchzuführen. Wir empfehlen, sich dabei auf folgende fünf Schritte zu konzentrieren:

1. **Prüfen Sie das aktuelle Potenzial Ihrer HPC-Infrastruktur.** Bewerten Sie Ihre Rechen-, (Arbeits-)Speicher und E/A-Ressourcen und finden Sie heraus, an welchen Stellen Sie eventuell investieren müssen, um Ihre Ressourcen für den Einsatz von KI zu optimieren. Am besten laufen HPC- und KI-Workloads auf hochleistungsfähigen Rechenplattformen wie dem Intel® Xeon® Platinum 9242 Prozessor, der eine durchschnittliche Leistungsverbesserung um das bis zu 3,7-fache[1] bei HPC-Benchmarks im Vergleich zu einem drei Jahre alten System erzielte. Bei KI-Workloads erzielt der Intel® Xeon® Platinum 8280 Prozessor im Vergleich zu Prozessoren der vorherigen Generation eine 14-mal[2] so hohe Deep-Learning-Inferenz durch seinen Intel® Deep Learning Boost. Während der Intel® Xeon® Platinum 9282 Prozessor im Vergleich zu Prozessoren der vorherigen Generation eine 30-mal[3] so hohe Deep-Learning-Inferenz durch seinen Intel® Deep Learning Boost und die Intel® Optimierung für Caffe erzielt hat. Zu anderen wichtigen Anforderungen gehören Verbindungen mit geringer Latenz wie bei der Intel® Omni-Path-Architektur (Intel® OPA) und mit ausreichend

Speicher für große Datenmengen wie die Intel® Optane™ Solid-State-Laufwerke (SSDs) und das neue Intel® Optane™ DC Persistent Memory. Wenn Sie gerade erst begonnen haben, Ihren KI-Bedarf zu prüfen: Stellen Sie sicher, dass Ihnen die Erwartungen der unterschiedlichen Stakeholder im Unternehmen klar sind, gleichen Sie sie mit ihren technischen Ressourcen ab und identifizieren Sie eventuelle Lücken.

2. **Verschaffen Sie sich einen Überblick über die unterschiedlichen KI-Frameworks und -Bibliotheken** und finden Sie heraus, welche Ihren Anforderungen am besten entsprechen. Es gibt bereits eine Reihe von Frameworks (wie Tensorflow*, Caffe* und MXNet*), mit denen Apps für viele gängige KI-Anwendungsfälle erstellt werden können, einschließlich Bilderkennung, Übersetzung, Empfehlungs-Engines und Generative Adversarial Networks.
3. **Stellen Sie sicher, dass die ausgewählten KI-Frameworks optimiert sind** für Ihre aktuelle HPC-Infrastruktur (indem Sie die neuesten optimierten Framework-Versionen herunterladen). Das ist entscheidend, denn so können Sie sicher sein, dass Sie mit Ihren KI-Anwendungen jederzeit die größtmögliche Skalierbarkeit, Effizienz und Leistung erreichen.
4. Vielleicht entscheiden Sie sich auch dafür, Ihre eigenen Algorithmen zu entwickeln. In diesem Fall sollten Sie **sich darauf konzentrieren, diese von Anfang an für Ihre bestehende Intel®-Architekturumgebung zu optimieren**. An diesem Punkt sollten Sie noch nicht massiv in neue Hardware wie Beschleuniger und GPUs investieren müssen. Wenn Sie Ihre Algorithmen mit gebrauchsfertigen Tools in der Cloud testen wollen, achten Sie auch hier auf Ressourcen, die auf Ihre Infrastruktur abgestimmt sind. Beispielsweise hat Intel in enger Zusammenarbeit mit Amazon Web Services das Algorithmentraining in der Cloud optimiert. Bei der Softwareentwicklung sollten Sie die neuesten Tools nutzen, um für einen möglichst schnellen und reibungslosen Prozess zu sorgen.
5. **Planen Sie Ihre Workloads sorgfältig.** Überlegen Sie sich genau, zu welchen Teilen Ihre KI-Maßnahmen auf Training und zu welchen Teilen sie auf Inferenz abzielen soll. Dafür sind jeweils andere Kompetenzen notwendig.

Framework-Optimierungen

Intel setzt alles daran, dass seine neuesten, auf skalierbaren Intel® Xeon® Prozessoren aufsetzenden Rechenplattformen für die gängigsten KI-Frameworks optimiert sind. Zu diesen Optimierungen gehören z. B.:

- Intel® Optimierung für TensorFlow*: Dieses Python*-basierte Deep Learning Framework ist ausgelegt auf einfache Benutzung und Erweiterbarkeit auf modernen Deep Neural Networks.
- Intel® Optimierung für Caffe*: Diese optimierte Version von Caffe* ist eines der beliebtesten Frameworks für Bilderkennung.
- MXNet*: Dieses frei zugängliche Deep Learning Framework verfügt über integrierten Support für die Intel® Math Kernel Library (Intel® MKL).

Bereit für Skalierung und Wachstum

Sobald Sie Ihren ersten experimentellen KI-Workload erfolgreich implementiert und ausgeführt haben, besteht die nächste Herausforderung darin, ihn zu skalieren und im gesamten Unternehmen auszurollen. Im Hinblick auf dieses Ziel sollte die Anwendung darauf ausgerichtet sein, die Funktionen Ihres HPC-Systems möglichst umfänglich zu nutzen, auf allen Kernen der CPU ausgeführt zu werden und über mehrere Knoten in Ihrem HPC-Cluster zu skalieren. Nicht zuletzt sollte sie auch hierfür auf den neuesten Framework-Versionen aufsetzen. Der Grad, den Sie bei der Erreichung dieser drei Zielsetzungen erreichen, sowie die für die Trainingsmodelle geschätzte Dauer und Frequenz sind wertvolle Indikatoren: Dafür, wie effektiv und effizient Sie skalieren können, und/oder dafür, an welcher Stelle Sie in Rechenressourcen investieren müssen, bevor Sie loslegen.

Für diejenigen, die so schnell wie möglich starten wollen, hat Intel zusammen mit anderen Branchenführern die Intel® Select Solutions für HPC entwickelt. Diese Suite enthält Technologien für eine schnell einsatzbereite Infrastruktur, die für Analyse-Cluster und HPC-Anwendungen optimiert ist. Damit kann die Zeit bis zum Durchbruch, bis zum Gewinnen verwertbarer Erkenntnisse und bis zur Umsetzung in einem neuen Produktdesign erheblich verkürzt werden.

Egal, mit welchem KI-Workload Sie beginnen – die vollständige Konvergenz mit Ihrer HPC-Umgebung wird einiges an Zeit kosten. Auf Ihrem Weg zum Ziel werden sich neue Herausforderungen, aber auch neue Chance auftun, denn die Möglichkeiten von HPC und KI entwickeln sich ständig weiter. Solange Sie Ihre Unternehmensziele fest im Blick behalten und Ihre KI-Maßnahmen konsequent danach ausrichten, sind Sie optimal aufgestellt für langfristigen Erfolg.

Weitere Informationen

- Tutorial: Mit TensorFlow* Deep Learning trainieren und testen
- Solution Brief: Leistung für HPC-Plattformen
- Solution Brief: Ausführen von KI- und Analyseanwendungen
- Solution Brief: Leistung für HPC-Plattformen

Finden Sie die passende Lösung für Ihr Unternehmen. Wenden Sie sich an Ihren Ansprechpartner bei Intel oder besuchen Sie **www.intel.de/hpc**

In Leistungstests verwendete Software und Workloads können speziell für die Leistungseigenschaften von Intel® Mikroprozessoren optimiert worden sein.

Leistungstests wie SYSmark* und MobileMark* werden mit spezifischen Computersystemen, Komponenten, Softwareprogrammen, Operationen und Funktionen durchgeführt. Jede Veränderung bei einem dieser Faktoren kann abweichende Ergebnisse zur Folge haben. Als Unterstützung für eine umfassende Bewertung Ihrer geplanten Anschaffung sollten Sie noch andere Informationen und Leistungstests heranziehen – auch im Hinblick auf die Leistung des betreffenden Produkts in Verbindung mit anderen Produkten. Ausführlichere Informationen finden Sie unter www.intel.de/benchmarks.

Die Leistungsergebnisse basieren auf Tests, die zu dem in den Konfigurationen angegebenem Datum durchgeführt wurden, und spiegeln möglicherweise nicht alle öffentlich erhältlichen Sicherheitsupdates wider. Weitere Einzelheiten finden Sie in den veröffentlichten Konfigurationsdaten. Kein Produkt und keine Komponente bietet absolute Sicherheit.

[1] Durchschnittliche Zunahme um das 3,7-Fache mit dem Intel® Xeon® Platinum 9242 Prozessor im Vergleich zu einem 3 Jahre alten Server: Durchschnittliches geometrisches Mittel von STREAM, HPCG, HPL, WRF, OpenFOAM, LS-Dyna, VASP, NAMD, LAMMPS, Black Scholes und Monte Carlo. Der jeweilige Workload kann variieren. Intel® Xeon® E5-2697 v4 Prozessor: Intel-Referenzplattform mit 2S Intel® Xeon® E5-2697 v4 Prozessoren (2,3 GHz, 18C), 8 x 16 GB DDR4-2400, 1 SSD, Cluster-Dateisystem: Panasas (124 TB Speicher), Firmware v6.3.3.a und OPA-basiertem IEEL Lustre, BIOS: SE5C610.86B.01.01.0027.071020182329, Mikrocode: 0xb00002e, Oracle Linux Server Version 7.6 (kompatibel mit RHEL 7.6) auf einem 7.5-Kernel mit ksplice zur Sicherheitskorrektur, Kernel: 3.10.0-862.14.4.el7.crt1. x86_64, OFED-Stack: OFED OPA 10.8 unter RH7. 5 mit Lustre v2.10.4, HBA: 1-Port-PCIe-x16-Adapter für Intel® OPA, Switch: Intel OPA Edge Switch Serie 100 mit 48 Ports. Stream OMP 5.10, Triad, HT=EIN, Turbo=AUS, 1 Thread pro Kern, Ergebnis: 128,36. HPCG, Binärcode enthalten, MKL 2019u1, HT=EIN, Turbo=AUS, 1 Thread pro Kern, Ergebnis: 23,78. HPL 2.1, HT=EIN, Turbo=AUS, 2 Threads pro Kern, Ergebnis: 1.204,64. WRF 3.9.1.1, conus-2.5km, HT=EIN, SMT=EIN, 1 Thread pro Kern, Ergebnis: 4,54. OpenFOAM 6.0, 42M_cell_motorbike, HT=EIN, Turbo=AUS, 1 Thread pro Kern, Ergebnis: 3.500. LS-Dyna 9.3-Explicit AVX2 Binärcode, 3car, HT=EIN, SMT=EIN, 1 Thread pro Kern, Ergebnis: 2.814. VASP 5.4.4, CuC, HT=EIN, Turbo=AUS, 1 Thread pro Kern, Ergebnis: 384,99. NAMD 2.13, apoa1, HT=EIN, Turbo=AUS, 2 Threads pro Kern, Ergebnis: 4,4. LAMMPS Version 12 Dez 2018, Water, HT=EIN, Turbo=EIN, 2 Threads pro Kern, Ergebnis: 54,72. Black Scholes, HT=EIN, Turbo=EIN, 2 Threads pro Kern, Ergebnis: 2.573,77. Monte Carlo, HT=EIN, Turbo=EIN, 2 Threads pro Kern, Ergebnis: 43,2. Intel® Xeon® 9242 Prozessor: Intel-Referenzplattform mit 2S Intel® Xeon® 9242 Prozessoren (2,2 GHz, 48C), 24 x 16 GB DDR4-2933, 1 SSD, Cluster-Dateisystem: 2.12.0-1 (Server) 2.11.0-14.1 (Client), BIOS: PLYXCRB1.86B.0572. D02.1901180818, Mikrocode: 0x4000017, CentOS 7.6, Kernel: 3.10.0-957.5.1.el7. x86_64, OFED-Stack: OFED OPA 10.8 unter RH7. 5 mit Lustre v2.10.4, HBA: 1-Port-PCIe-x16-Adapter für Intel® OPA, Switch: Intel OPA Edge Switch Serie 100 mit 48 Ports. STREAM OMP 5.10, Triad, HT=EIN, Turbo=AUS, 1 Thread pro Kern, Ergebnis: 407. HPCG, Binärcode enthalten, MKL 2019u1, HT=EIN, Turbo=AUS, 1 Thread pro Kern, Ergebnis: 81,91. HPL 2.1, HT=EIN, Turbo=AUS, 2 Threads pro Kern, Ergebnis: 5314. WRF 3.9.1.1, conus-2.5km, HT=EIN, SMT=EIN, 1 Thread pro Kern, Ergebnis: 1,44. OpenFOAM 6.0, 42M_cell_motorbike, HT=EIN, Turbo=AUS, 1 Thread pro Kern, Ergebnis: 1.106. LS-Dyna 9.3-Explicit AVX2 Binärcode, 3car, HT=EIN, SMT=EIN, 1 Thread pro Kern, Ergebnis: 768. VASP 5.4.4, CuC, HT=EIN, Turbo=AUS, 1 Thread pro Kern, Ergebnis: 133,96. NAMD 2.13, apoa1, HT=EIN, Turbo=AUS, 2 Threads pro Kern, Ergebnis: 19,9. LAMMPS Version 12 Dez 2018, Water, HT=EIN, Turbo=EIN, 2 Threads pro Kern, Ergebnis: 276,1. Black Scholes, HT=EIN, Turbo=EIN, 2 Threads pro Kern, Ergebnis: 9.044,32. Monte Carlo, HT=EIN, Turbo=EIN, 2 Threads pro Kern, Ergebnis: 227,62.

Haftungsausschluss von OpenFOAM: Dieses Angebot wurde nicht von OpenCFD Limited, Hersteller und Distributor der OpenFOAM-Software über www.openfoam.com und Besitzer der Marken OPENFOAM® und OpenCFD®, genehmigt oder empfohlen. Testdatum: 15. März 2019.

[2] Die Leistungsergebnisse beruhen auf Tests vom 11.7.2017(1-mal), 8.11.2018 (5,7-mal), 20.2.2019 (14-mal) und 26.2.2019 (30-mal) und spiegeln möglicherweise nicht alle öffentlich verfügbaren Sicherheitsupdates wider. 14-fache Inferenzdurchsatzverbesserung mit Intel® Xeon® Platinum Prozessor 8280 mit Intel® DL Boost: Von Intel getestet. Stand: 20.2.2019. Zweiprozessorsystem mit Intel® Xeon® Platinum Prozessor 8280, 28 Kerne, HT=EIN, Turbo=EIN, insgesamt 384 GB Arbeitsspeicher (12 Steckplätze, je 32 GB, 2933 MHz), BIOS: SE5C620.86B.0D.01.0271.120720180605 (uCode: 0x200004d), Ubuntu 18.04.1 LTS, Kernel 4.15.0-45-generic, SSD: 1 x INTEL SSDSC2BA80 (sda) – SSD mit 745,2 GB, INTEL SSDPE2KX040T7 (nvme1n1) – SSD mit 3,7 TB, Deep Learning Framework: Intel® Optimierungen für Caffe*, Version: 1.1.3 (Commit-Hash: 7010334f159da247db3fe3a9d96a3116ca06b09a), ICC-Version 18.0.1, MKL-DNN-Version: 0.17 (Commit-Hash: 830a10059a018cd2634d94195140cf2d8790a75a, Modell: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS = 64, syntheticData, 4 Instanzen/Zweiprozessorsystem, Datentyp: INT8 im Vergleich zu Test durch Intel am 11. Juli 2017: Zweiprozessorsystem mit Intel® Xeon® Platinum Prozessor 8180 (2,50 GHz, 28 Kerne), HT deaktiviert, Turbo deaktiviert, Scaling-Governor festgelegt auf „Performance“ über intel_pstate-Treiber, 384 GB DDR4-2666-ECC-RAM. CentOS Linux, Release 7.3.1611 (Kern), Linux-Kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD der Produktreihe DC S3700 (800 GB, 2,5", 6-GBit/s-SATA, 25-nm-Technik, MLC).

Leistung gemessen mit: Umgebungsvariable: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPUFreq festgelegt mit: cpupower frequency-set -d 2,5 G -u 3,8 G -g Performance. Caffe: (http://github.com/intel/caffe/), Version f96b759f71b2281835f690af267158b82b150b5c. Inferenz gemessen mit „caffe time --forward_only“-Befehl, Training gemessen mit „caffe time“-Befehl. Für die „ConvNet“-Topologien wurde ein synthetischer Datensatz verwendet. Für andere Topologien wurden Daten im lokalen Datenspeicher gespeichert und vor dem Training im Systemspeicher zwischengespeichert. Topologie-Spezifikationen von https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50). Intel C++ Compiler, Version 17.0.2 20170213, Intel MKL Small Libraries, Version 2018.0.20170425. Caffe ausgeführt mit „numactl -l“.

[3] 30-fache Inferenzdurchsatzverbesserung mit Intel® Xeon® Platinum Prozessor 9282 mit Intel® DL Boost: Von Intel getestet. Stand: 26.2.2019. Plattform: Dual-Sockel „Dragon Rock“ mit Intel® Xeon® Platinum 9282 Prozessor (56 Kerne pro Prozessor), HT aktiviert, Turbo aktiviert, insgesamt 768 GB Arbeitsspeicher (24 Steckplätze, je 32 GB, 2933 MHz), BIOS: SE5C620.86B.0D. 01.0241.112020180249, CentOS 7 Kernel 3.10.0-957.5.1.el7. x86_64, Deep Learning Framework: Intel® Optimierung für Caffe*, Version: https://github.com/intel/caffe d554cbf1, ICC 2019.2.187, MKL-DNN-Version: 0.17 (Commit-Hash: 830a10059a018cd-2634d94195140cf2d8790a75a, Modell https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS = 64, keine Datenebene – syntheticData: 3x224x224, 56 Instanzen / Dual-Sockel, Datentyp: INT8; Vergleich mit: Von Intel getestet am 11. Juli 2017: Dual-Sockel Intel® Xeon® Platinum 8180 Prozessor (2,50 GHz, 28 Kerne), HT=EIN, Turbo=AUS, Scaling-Governor festgelegt auf „Performance“ über intel_pstate-Treiber, 384 GB DDR4-2666-ECC-RAM. CentOS Linux, Release 7.3.1611 (Kern), Linux-Kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD der Produktreihe DC S3700 (800 GB, 2,5", 6-GBit/s-SATA, 25-nm-Technik, MLC). Leistung gemessen mit: Umgebungsvariablen: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPUFreq festgelegt mit: cpupower frequency-set -d 2,5G -u 3,8G -g Performance. Caffe: (http://github.com/intel/caffe/), Version f96b759f71b2281835f690af267158b82b150b5c. Inferenz gemessen mit „caffe time --forward_only“-Befehl, Training gemessen mit „caffe time“-Befehl. Für die „ConvNet“-Topologien wurde ein synthetischer Datensatz verwendet. Für andere Topologien wurden Daten im lokalen Datenspeicher gespeichert und vor dem Training im Systemspeicher zwischengespeichert. Topologie-Spezifikationen von https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50). Intel C++ Compiler, Version 17.0.2 20170213, Intel MKL Small Libraries, Version 2018.0.20170425. Caffe ausgeführt mit „numactl -l“.

Lösung bereitgestellt von:

Alle hierin gemachten Angaben können sich jederzeit ohne besondere Mitteilung ändern. Wenden Sie sich an Ihren Ansprechpartner bei Intel, um die neuesten Produktspezifikationen und Roadmaps zu erhalten.

Durch Technologien von Intel ermöglichte Funktionsmerkmale und Vorteile hängen von der Systemkonfiguration ab und können entsprechend geeignete Hardware, Software oder die Aktivierung von Diensten erfordern. Die Leistung kann je nach Systemkonfiguration unterschiedlich ausfallen. Kein Computersystem bietet absolute Sicherheit. Informieren Sie sich beim Systemhersteller oder Fachhändler oder auf intel.de.

Intel, das Intel Logo, Xeon und Optane sind Marken der Intel Corporation oder ihrer Tochtergesellschaften in den USA und/oder anderen Ländern.

*Andere Marken oder Produktnamen sind Eigentum der jeweiligen Inhaber.

© Intel Corporation

0519/JO/CAT/PDF 338068-002DE